Investigating the use of Modulation Spectral Features within an I-vector Framework for Far-Field Automatic Speaker Verification

Anderson R. Avila, Francisco J. Fraga
CECS-UFABC, Universidade Federal do ABC
Santo André, São Paulo, Brazil

Milton Sarria-Paja, Tiago H. Falk
INRS-EMT, University of Quebec
Montreal, Quebec, Canada

***Abstract*—It is known that channel variability compromises automatic speaker recognition accuracy. However, little attention has been given so far to the detrimental effects encountered under reverberant environments. In this paper, we focus on the issue of automatic speaker verification (ASV) under several levels of room reverberation. Alternative auditory inspired features are explored. Specifically, we investigate whether the performance of the so-called modulation spectral features (MSFs) can overcome the well-known mel-frequency cepstral coefficients (MFCCs). Experiments were conducted with an ASV system based on the state-of-the-art i-vector. The main contribution of this paper is to verify if MSFs combined with i-vectors are able to present the same performance encountered in the literature regarding speech recognition and speaker identification systems in reverberant environment.**

I. INTRODUCTION

The acoustic signal captured in an enclosed environment is subject of reverberation due to the reflection on walls, floor and ceilings. This effect can degrade speech quality, affecting not only human speech perception but also the performance of many applications, such as those related to hands-free communication. Although speaker verification systems can achieve high accuracies under matched conditions (i.e., when testing and training data are obtained under the same circumstances), it is still a great challenge to maintain the same performance for real world applications where channel variability is inevitable.

Many efforts have been made in order to mitigate channel effects. Feature compensation techniques such as cepstral mean subtraction (CMS) [1] and RASTA [2] are meant to remove distortions artifacts from the feature vector before speaker enrollment takes place. Another formulation, related to the decision process, aims at reducing score variability through score normalization techniques such as Hnorm and Tnorm. Application examples and details on these methods can be found in [3]. A third approach that deals with channel effects regards modelling speakers considering session variability [4]. Joint Factor Analysis (JFA) has offered interesting performance by modelling separately inter-speaker and inter-session variability assuming that speaker factors remain the same for different recordings while channel factors vary [5]. More recently, an extension of JFA has been proposed in [6]. The new low-dimensional space, named total variability or i-vector, has been demonstrating state-of-the-art results and is motivated by the fact that channel factors also contains information about the speaker identity [7].

Modulation spectral features (MSFs) have been considered as an alternative to maintain performance in reverberant environments. The experiments carried out in [8] showed that, although MSFs offered lower performance when clean speech was used during the training and testing phase, they outperformed the results obtained by using other auditory inspired features, such as Perceptual Linear Prediction (PLP), when room reverberation was taken into account. Authors in [9] demonstrated the benefit of using MSFs for speech recognition and argue that their robustness against reverberant environments occurs because amplitude modulations are less susceptible to distortions caused by reverberation when compared to the fine structure of speech signals.

In [10] authors have shown that modulation bands ranging between 3 and 15 Hz are robust to increasing levels of reverberation time (T_{60}). They compared the performance of a speaker identification system based on MFCCs to a system based on MSFs. Results showed that the speaker identification system proposed, based on MFCCs, offered better performance for low levels of room reverberation (i.e., T_{60} up to 0.4s), with its performance decaying severely for higher levels of room reverberation. Their system, based on MSFs, outperformed MFCCs for reverberation time higher than 0.4s. In the experiments described herein, we made a similar comparison, but now considering a different dataset submitted to different environmental conditions and, most importantly, using the burgeoning total variability (i-vector) paradigm instead of the traditional GMM-UBM (Gaussian mixture model - Universal background model) framework.

This paper is organized as follows. Section 2 gives an overview on speaker verification systems (ASV) based on i-vector and describes the feature extraction steps and also the decision process used in this work. Section 3 discusses reverberation, presents the database used in this paper and also describes the experimental setup. The results and discussion are presented in section 4. Section 5 concludes the paper.

II. ASV SYSTEM

As described in Fig. 1, ASV systems are composed by a front-end responsible for the feature extraction, a modelling method used for speaker enrollment and a decision process usually based on likelihood score for GMM-UBM and on cosine similarity for i-vector. Such systems aim to solve the following question: *Is the speaker who he claims to be?* In

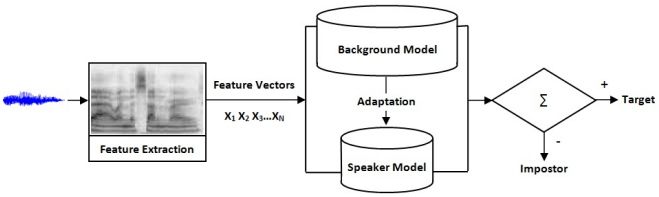


Fig. 1: Block diagram of automatic speaker verification.

other words, given an utterance and a claimed identity, verify if the claimer is the genuine speaker or an impostor. In this work, we consider a text-independent speaker verification using the i-vector framework, based on the MSR Identity Toolbox [11].

A. Feature extraction

1) Mel-Frequency Cespstral Coefficients (MFCC): As pre-processing steps, speech signals were downsampled to 8 kHz, pre-emphasized and normalized at -26 dBov (dB overload) and framed at every 10-ms using a 20-ms Hamming window. After applying the FFT, a set of 24 triangular bandpass filters obeying the Mel scale were used before the discrete cosine transform (DCT) [12], according to the following formula:

$$x_n = \sum_{m=1}^{M} [Y_m] cos \left[\frac{\pi n}{M} \left(m - \frac{1}{2} \right) \right], n = 1, 2, 3, ..., P, \quad (1)$$

where x_n is the n^{th} mel-cepstrum coefficient and Y_m refers to the log-energy of the m^{th} filter output. The set of P coefficients forms the MFCC feature vector, $\overrightarrow{x} = \{x_1, x_2, x_3, ..., x_p\}$.

In our experiments, we considered $M = 24$ and $P = 19$, which resulted in 19 cepstral coefficients plus the global log-energy. Delta and double-delta coefficients were then computed from the MFCCs and log-energy features, resulting in a 60-dimensional feature vector.

2) Modulation Spectral Features (MSF): As depicted in Figure 2, five signal processing steps were involved in our computation of modulation spectral features. During the pre-processing step, speech signals were downsampled either to 8 kHz or 16 kHz, as it will be observed in the experiments we performed. Also, an energy-thresholding voice activity detection (VAD) is used followed by an energy normalization to -26 dBov (dB overload).

Next, the speech signal passes through a bank of 23 critical-band gammatone filters, emulating the process performed by the cochlea. The first filter in the filterbank is centred at 125 Hz and the last one close to half the sampling rate (i.e. approximately 3.5 kHz and 7 kHz, respectively for 8 kHz and 16 kHz as sampling rates) [10]. Filter bandwidths are defined by the equivalent rectangular bandwidth (ERB), which is given by

$$ERB_j = \frac{f_j}{Q_{\text{ear}}} + B_{\text{min}} \quad (2)$$

where f_j represent the center frequency of the j-th filter and Q_{ear} and B_{min} are, respectively, set to 9.265 and 24.7 [10]. The signal $s_j(n)$, originated by j-th filter, is then used to compute the temporal envelope $e_j(n)$ through Hilbert Transform, defined as

$$e_j(n) = \sqrt{s_j(n)^2 + \mathcal{H}\{s_j(n)\}^2} \quad (3)$$

Following, temporal envelopes $e_j(n)$ are multiplied every 32-ms by a 256-ms Hamming window. The discrete Fourier transform of the temporal envelope is computed in order to obtain the modulation spectrum. Dropping the variable n for simplicity, we have

$$E_j(m; f) = |\mathcal{F}(e_j(n))| \quad (4)$$

where m represents the m-th frame obtained after every Hamming window multiplication and f designates modulation frequency. Lastly, an auditory-inspired modulation filterbank is used to represent the modulation frequencies into eight bands, denoted as $E_{j,1}, ..., E_{j,k}, ..., E_{j,8}$, where j represents the j-th critical band signal and k is the k-th modulation filterbank output.

B. Total Variability

In JFA [5], [13], given an utterance, the speaker and channel components are represented by a supervector *M*, defined as follows

$$M = m + Vy + Ux + Dz \quad (5)$$

where m is the speaker and session-independent supervector, being V the eigenvoice matrix and D the diagonal residual, both representing the speaker subspace ($s = m + Vy + Dz$). U is the eigensession matrix and it represents the session subspace ($c = Ux$).

Instead of two distinct spaces for modelling speaker and channel variability, as briefly described above, authors in [7] propose the use of a simple space, referred to as total variability space. The reasoning behind this new approach relies on the fact that channel factors estimated by JFA contains information about speakers too, as shown in the experiments performed by them [7]. Hence, for a given utterance, both speaker and session components represented by (7) can be rewritten as

$$M = m + Tw \quad (6)$$

where m is the mean supervector extracted from the universal background model, T corresponds to a retangular low-rank matrix and w is a random vector with normal distribution. The so-called hidden variable w contains the component factors and is referred to as the identity vector (i.e., the i-vector) [7].

Within the i-vector framework, which is depicted in Fig. 3, the decision process in the total variability space consists basically in computing the similarity between the target speaker factors and the test speaker factors. In our experiments, we

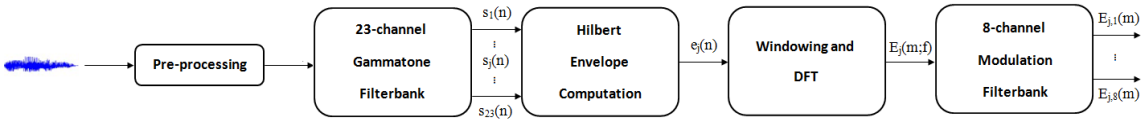


Fig. 2: Block diagram describing the signal processing steps used to extract the modulation spectral features.

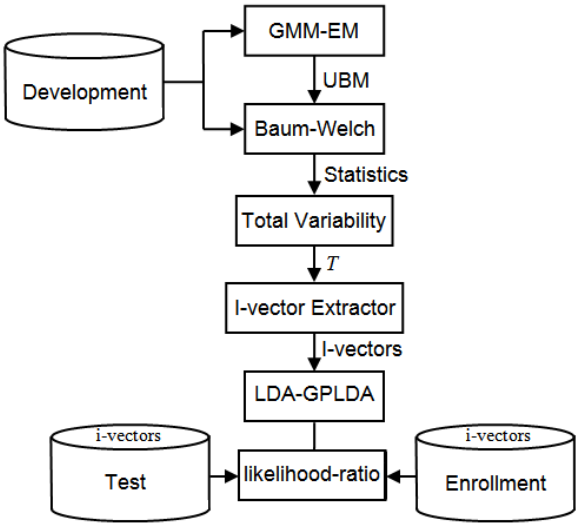


Fig. 3: Block diagram of the i-vector framework. Adapted from [11].

used 50[1] total factors defined by the total variability matrix T which was obtained from the combination of the UBM and the Baum-Welch statistics [7][14]. The i-vectors are then submitted to Linear Discriminant Analysis (LDA) which seeks to maximizes the between class variance and minimize the within class variance. The total factors are reduced to a 35-dimensional vector. Following the LDA, EM algorithm is used to obtain a Gaussian Probabilistic Linear Discriminant Analysis (PLDA) model [15]. The decision process is base on the log-likelihood and is defined [16] as

$$score = log \frac{p\left(x_1, x_2 \mid H_0\right)}{p\left(x_1 \mid H_1\right) p\left(x_2 \mid H_1\right)} \tag{7}$$

where x_1 and x_2 are the i-vectors involved in a trial and are represented by PLDA hyperparameters. H_0 and H_1 represent the hypothesis described in section 2.

III. EXPERIMENTAL SETUP

In this section, we present the details of the experiments conducted in order to compare the performance of both features (i.e. mel-frequency cepstral coefficients and modulation spectral features) under different T_{60}. Their performances are analyzed under the i-vector framework.

[1]We have not found any significant improvements by considering higher values for the total factors.

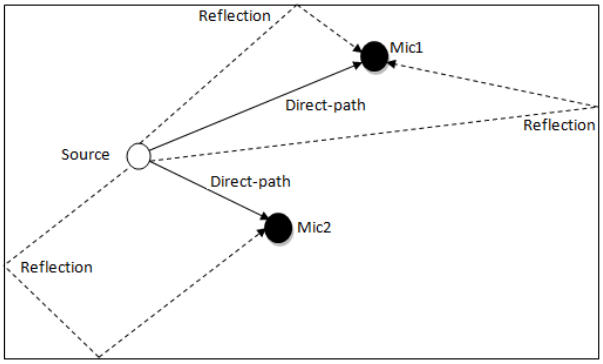


Fig. 4: Room reverberation scheme.

A. Reverberation

In reverberant environments the speech signal propagates from a source (i.e., a speaker) to one or more microphones as depicted in Fig. 4. The direct-path is the acoustic propagation path between the speaker and the microphones with no reflections. A number of delayed and attenuated versions of the original signal reaches the microphones superimposing the original signal [17]. Assuming a linear time-invariant system (LTI), the reverberant signal $r_m(n)$ captured by the m-th microphone can be modelled by the convolution of the source speech signal $s(n)$ and the acoustic room impulse response $h_m(n)$, as shown by

$$r_m(n) = s(n) * h_m(n) + v_m(n), \quad m = 1, ..., M. \tag{8}$$

where m denotes the m-th microphone placed in a reverberant room and $v_m(n)$ represents the environment noise received by the m-th microphone. In this work, the reverberant speech were generated by convolving recorded room impulse responses with clean speech signals.

B. Database Description

The clean speech part of Chains corpus [18] was used in our experiments. It features recordings of 36 subjects, including male and female with different accents. Six different styles are provided: SOLO, SYNCHRONOUS, RETELL, RSI, WHISPER, FAST. The first three speaking styles were recorded in a professional studio and the speakers were placed in a booth while recording. The last three styles were recorded in a quiet office environment. For this study, we used only SOLO (i.e., normal speech). Each speaker read 37 prepared text which generated 37 distinct speech files. The content read was always the same independently of the type of vocal effort used by the subject. The normal style was obtained with the subjects reading each text at a comfortable rate. During the

enrollment phase, the four first speech files of each speaker were used for training, which added up to roughly 120 seconds. The remaining 33 speech files were used for testing and it provided approximately 60 seconds of speech.

Room impulse response (RIR) were obtained in a professional chamber (length 7 m, width 9 m x height 3.5 m). Measurements were conducted using four powered loudspeakers positioned in the lower corners of the room. The RT_{60} was measured and recorded by a microphone placed in the center of the room and absorbers were used to manage the reverberation level. RT_{60} values collected include: 0.39, 0.45, 0.55, 0.63, 0.75, 0.87, 1.04, 1.15, 1.41 and 2.03.

Throughout the experiments conducted here, the trials involved clean speech and also reverberant speech obtained by convolving the testing set (i.e., the last 33 sentences of the Chain Corpus for each speaker) with different levels of reverberation, characterized by all the RIR previously listed.

C. Three Modulation Bands

Three modulation channels, $E_{j,k}$ $(k = 1, 2, 3)$ were considered in the first experiment. Since each one of these channels operates over all outputs of a 23-channel gammatone filterbank, a 69-dimensional energy vector is generated per frame. Each frame is normalized by the maximum modulation energy among modulation bands $k = 1, 2, 3$. Following the normalization, principal component analysis (PCA) is applied in order to maximize the variance of the data by projecting it onto a 23-dimensional space.

For enrollment, as it occurred in the previous experiments, the first four speech files of each speaker presented in the Chains corpus were used, leading to approximately 120 seconds of data. Only clean speech was used for training. In the verification phase, the remaining 33 speech files were used for testing. Clean speech files were corrupted with different reverberation levels during testing. The same procedure will apply for the next two experimental setups.

D. Four Modulation Bands

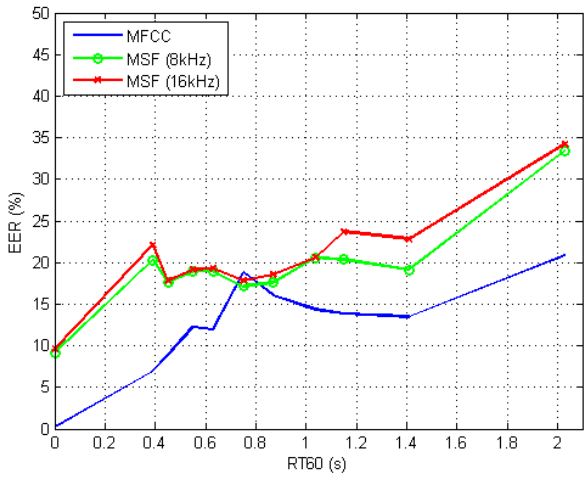
In the second approach, four modulation channels were considered, which led to a 92-dimensional energy vector per frame. As has happened previously, each frame is normalized by the maximum modulation energy among modulation bands $k = 1, 2, 3, 4$. After, PCA was applied, projecting the data onto a 46-dimensional space.

E. Five Modulation Bands

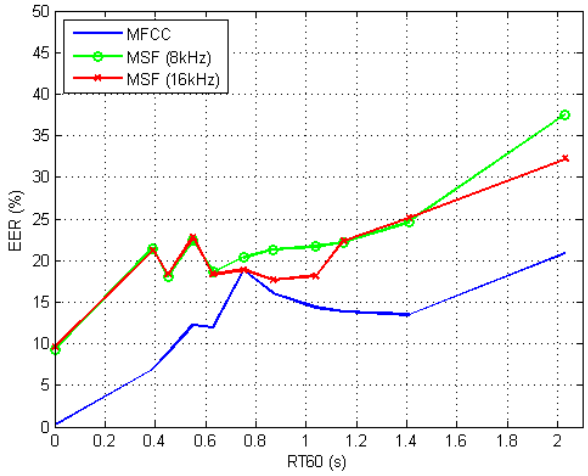
Finally, we investigated the use of five modulation channels, which led to a 115-dimensional energy vector per frame. As usual, each frame is normalized by the maximum modulation energy among modulation bands $k = 1, 2, 3, 4, 5$. Principal component analysis (PCA) is then applied, which projected the data onto a 60-dimensional space.

IV. RESULTS AND DISCUSSION

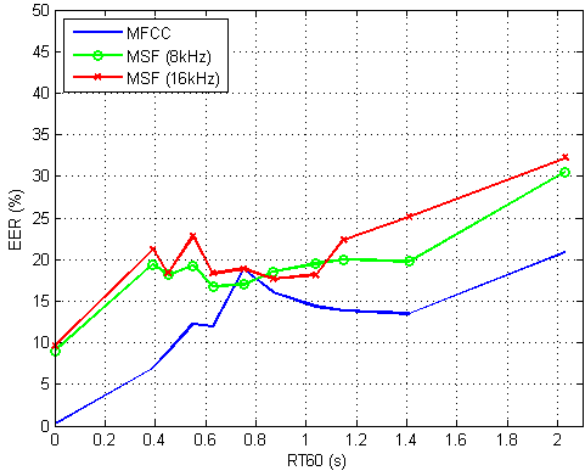
In Figure 5, results of the aforementioned experiments are presented and compared to the results obtained with the baseline system, which consisted of using MFCCs as features.



(a) 3 MSF bands



(b) 4 MSF bands



(c) 5 MSF bands

Fig. 5: Equal error rate vs T_{60} using MFCC and MSF.

Setup	min	max	$\overline{x}$	σ^2
MFCC	0.18	20.93	11.55	32.54
MSFs / 3 bands / 8 kHz	9.18	33.40	19.36	31.59
MSFs / 3 bands / 16 kHz	9.62	34.25	20.50	34.74
MSFs / 4 bands / 8 kHz	9.26	37.54	21.59	43.98
MSFs / 4 bands / 16 kHz	9.62	32.22	20.42	31.17
MSFs / 5 bands / 8 kHz	9.05	30.55	18.90	24.45
MSFs / 5 bands / 16 kHz	9.91	44.15	26.80	79.21

TABLE I: I-vector performance for different number of modulation bands and two different sampling rate.

Clean speech was used during the enrollment process and reverberant speech for testing. MFCCs offered the best results compared to the MSFs, as can be also seen in Table I. We observed only a slightly difference between the two sampling rates considered for MSFs, with 8 kHz offering better results, except when 4 modulation bands are used.

Overall, MSFs failed to outperform MFCCs, on the contrary to what have been shown in other studies. However, those studies were not directly related to speaker recognition [8][9] nor to speaker verification [10] and therefore universal background models were not involved in their experiments, which may be one possible explanation for MSFs not being able to outperform MFCCs. Notice in Figure 5 that when compared to the baseline results (i.e., the results from MFCCs), as the level of reverberation time increases, the plots from our experiments offer similar slopes, which suggest that MSFs were equally affected by the increase of T_{60}.

Notwithstanding, MSFs performance seems stable over the 0.4-1.4 range, which is aligned with results obtained in [10], where authors had considered a smaller range (0-1s) and thus there is a possibility that their system based on MSFs would degrade with RT60 higher than 1s, as has happened to our results, although we cannot state it for sure. Actually, our experiments did not reveal any advantage on the use of MSFs instead of the well-known MFCCs, as was observed on the experiments performed in [10]. Beside the fact we used different datasets and room impulse responses, authors in [10] investigated speaker identification and not speaker verification as was done in this study. Another possibility to explain these opposite outcomes may be due to the replacement of traditional UBM-GMM models by a state-of-the-art i-vector based ASV system, for i-vectors are intrinsically built to deal with intersession variability. Our hypothesis is that this upgrade from UBM-GMM to total variability framework has improved the overall system performance and the potential gain that could be provided using the MSFs alone has somehow been masked by this substitution.

V. CONCLUSION

In this paper, we evaluated the performance of an ASV system under the use of two distinct auditory inspired features. The classic mel-frequency cepstral coefficients and the so-called modulation spectral features were compared. Results showed that, for the speaker verification system considered in this work, MSFs are equally affected by reverberation time and have their performance degraded more than MFCCs as the level of T_{60} increases. Moreover, different number of modulation bands offered about the same results. Future work could consider the possibility of combining both features, in order to find out whether MSFs provide complementary information to MFCCs, i.e., would a system that combines the two modalities perform better than each modality alone.

ACKNOWLEDGEMENT

The authors acknowledge funds from the Centre for Advanced Systems and Technologies in Communications (SYTACOM) and the Emerging Leaders in the Americas Program (ELAP).

REFERENCES

[1] S. Furui, "Cepstral analysis technique for automatic speaker verification," *Trans. IEEE on Speech and Audio Processing*, vol. 29, no. 2, pp. 254–272, 1981.

[2] H. Hermansky, "Perceptual linear predictive (plp) analysis of speech," *Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, April 1990.

[3] F. B. et. al., "A tutorial on text-independent speaker verification," *EURASIP Journal of Applied Signal Processing*, vol. 2004, pp. 430–451, Jan. 2004.

[4] D. A. Reynolds, "Automatic speaker recognition using gaussian mixture speaker models," *The Lincoln Laboratory Journal*, vol. 8, pp. 173–192, November 1995.

[5] P. Kenny, "Joint factor analysis of speaker and session variability: Theory and algorithms," Tech. Rep., 2005.

[6] N. D. et. al., "Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification," in *Proc. Interspeech*, vol. 9, 2009, pp. 1559–1562.

[7] N. Dehak, R. Dehak, P. Kenny, P. Ouellet, and P. Dumouchel, "Front-end factor analysis for speaker verification," *Trans. IEEE on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[8] B. E. Kingsbury, N. Morgan, and S. Greenberg, "Robust speech recognition using the modulation spectrogram," *Speech Communication*, vol. 25, no. 13, pp. 117 – 132, 1998. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167639398000326

[9] N. Moritz, J. Anemuller, and B. Kollmeier, "Amplitude modulation spectrogram based features for robust speech recognition in noisy and reverberant environments," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, May 2011, pp. 5492–5495.

[10] T. Falk and W.-Y. Chan, "Modulation spectral features for robust far-field speaker identification," *Trans. IEEE on Audio, Speech, and Language Processing*, vol. 18, no. 1, pp. 90–100, 2010.

[11] S. Sadjadi, M. Slaney, and L. Heck, "MSR Identity Toolbox v1. 0: A MATLAB Toolbox for Speaker Recognition Research," *Speech and Language Processing Technical Committee Newsletter*, 2013.

[12] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *Trans. IEEE on Acoustics, Speech, and Signal Processing*, vol. 80, no. 4, pp. 357–366, August 1980.

[13] N. D. et. al., "Comparison between factor analysis and GMM support vector machines for speaker verification," in *Proc. IEEE-Odyssey of the Speaker and Language Recognition Workshop*, 2008.

[14] P. Kenny, "A small foot-print i-vector extractor," in *Proc. IEEE-Odyssey of the Speaker and Language Recognition Workshop*, 2012.

[15] S. Prince and J. Elder, in *Proc. IEEE 11th Int. Conf. on Computer Vision (ICCV)*.

[16] D. Garcia-Romero and C. Epsy-Winson, "Analysis of i-vector length normalization in speaker recognition systems," in *Proc. Interspeech*, 2011, pp. 249–252.

[17] P. Naylor and N. Gaubitch, Eds., *Speech Dereverberation*. Springer, 2010.

[18] F. C. et. al., "The chains speech corpus: Characterizing individual speakers," in *Proc. Int. Conf. on Speech and Computer (SPECOM)*, 2006, pp. 1–6.